How to Prevent the Continuous Damage of Noises to Model Training?

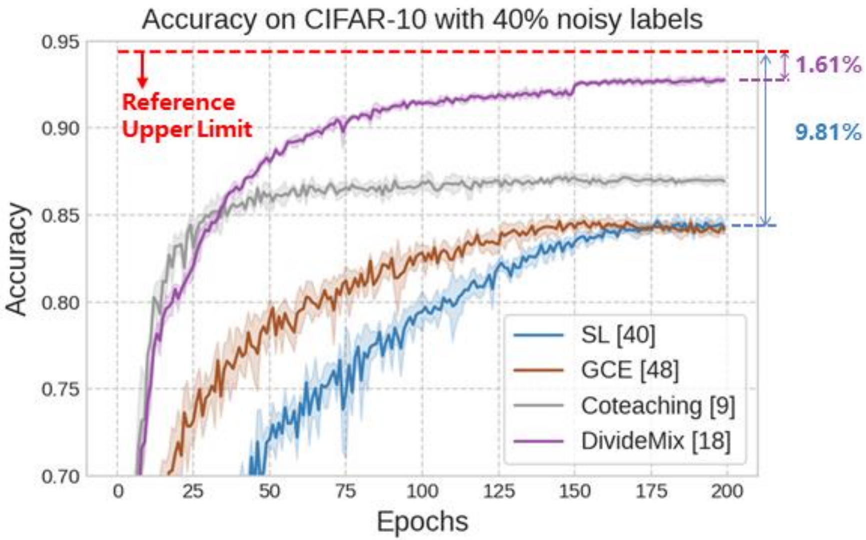
Xiaotian Yu[1], Yang Jiang[5], Tianqi Shi[5], Zunlei Feng[1,2], Yuexuan Wang[4], Mingli Song[1,2], Li Sun[3]*

[1]Zhejiang University,
[2]Shanghai Institute for Advanced Study of Zhejiang University,
[3]Ningbo Innovation Center Zhejiang University,
[4]University of Hong Kong, [5]Alibaba Group

Poster: THU-AM-365

Background

With existing methods, there are still large performance gaps between models trained with noisy samples and models trained with clean samples.



This phenomenon raises two questions:
How Noisy Labels Affect the Training and *Why Do Existing Methods Have Limited Effects*.

How Noisy Labels Affect the Training

a^l: The activated feature map of the l-th layer

m^l: The output feature map of the l-th layer

w^l: The kernel weight of the l-th convolution layer

$$m^l = a^{l-1} \otimes w^l$$

$$m^l_{i,j} = \sum_{i'} \sum_{j'} w^l_{i',j'} a^{l-1}_{is+i',js+j'}$$

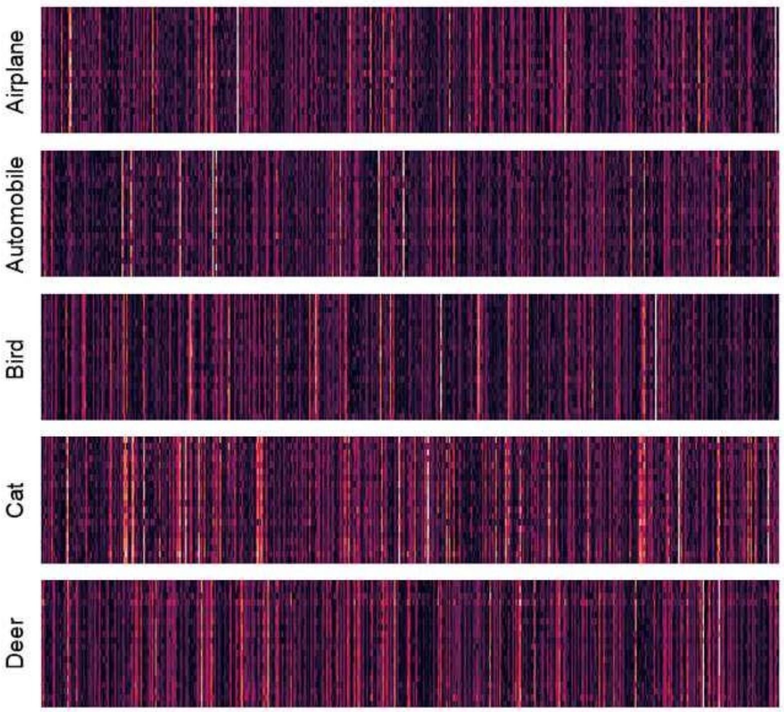
$$\frac{\partial \mathcal{L}(\tilde{y})}{\partial w^l_{i',j'}} = \sum_i \sum_j \frac{\partial \mathcal{L}(\tilde{y})}{\partial m^l_{i,j}} \frac{\partial m^l_{i,j}}{\partial w^l_{i',j'}}$$

$$= \sum_i \sum_j dil_s \left(\frac{\partial \mathcal{L}(\tilde{y})}{\partial m^l} \right)_{is,js} a^{l-1}_{i'+is,j'+js}$$

$$\frac{\partial \mathcal{L}(\tilde{y})}{\partial w^l} = a^{l-1} \otimes dil_s \left(\frac{\partial \mathcal{L}(\tilde{y})}{\partial m^l} \right)$$

$$\frac{\partial \mathcal{L}(\tilde{y})}{\partial m^l} = \sum_k \underbrace{\frac{\partial \mathcal{L}(\tilde{y})}{\partial z_k}}_{\text{Gradient Weight}} \overbrace{\frac{\partial z_k}{\partial m^l}}^{\text{Gradient Direction}}$$

The visualization of $\frac{\partial z_k}{\partial m^l}$:



The gradients $\frac{\partial z_k}{\partial m^l}$ of the same category have similar distributions, which can be regarded as gradient direction.

How Noisy Labels Affect the Training

a^l: The activated feature map of the l-th layer
m^l: The output feature map of the l-th layer
w^l: The kernel weight of the l-th convolution layer

$$m^l = a^{l-1} \otimes w^l$$

$$m^l_{i,j} = \sum_{i'} \sum_{j'} w^l_{i',j'} a^{l-1}_{is+i',js+j'}$$

$$\frac{\partial \mathcal{L}(\tilde{y})}{\partial w^l_{i',j'}} = \sum_i \sum_j \frac{\partial \mathcal{L}(\tilde{y})}{\partial m^l_{i,j}} \frac{\partial m^l_{i,j}}{\partial w^l_{i',j'}}$$

$$= \sum_i \sum_j dil_s \left(\frac{\partial \mathcal{L}(\tilde{y})}{\partial m^l} \right)_{is,js} a^{l-1}_{i'+is,j'+js}$$

$$\frac{\partial \mathcal{L}(\tilde{y})}{\partial w^l} = a^{l-1} \otimes dil_s \left(\frac{\partial \mathcal{L}(\tilde{y})}{\partial m^l} \right)$$

$$\frac{\partial \mathcal{L}(\tilde{y})}{\partial m^l} = \sum_k \underbrace{\frac{\partial \mathcal{L}(\tilde{y})}{\partial z_k}}_{\text{Gradient Weight}} \underbrace{\frac{\partial z_k}{\partial m^l}}_{\text{Gradient Direction}}$$

For CE loss, $\frac{\partial \mathcal{L}(\tilde{y})}{\partial z_k} = p_k - q_k \qquad (q_k = \mathbb{1}[\tilde{y} = k])$

$$\frac{\partial \mathcal{L}(y)}{\partial w^l} - \frac{\partial \mathcal{L}(\tilde{y})}{\partial w^l} = a^{l-1} \otimes dil_s \Bigg(\left(\frac{\partial \mathcal{L}(y)}{\partial z_y} - \frac{\partial \mathcal{L}(\tilde{y})}{\partial z_y} \right) \frac{\partial z_y}{\partial m^l}$$

$$+ \left(\frac{\partial \mathcal{L}(y)}{\partial z_{\tilde{y}}} - \frac{\partial \mathcal{L}(\tilde{y})}{\partial z_{\tilde{y}}} \right) \frac{\partial z_{\tilde{y}}}{\partial m^l} \Bigg)$$

$$= a^{l-1} \otimes dil_s \left(\frac{\partial z_{\tilde{y}}}{\partial m^l} - \frac{\partial z_y}{\partial m^l} \right).$$

The model trained with mislabeled samples generates gradient deviation, which will be accumulated and cause continuous damage. That is **how noisy labels affect the training.**

Why Do Existing Methods Have Limited Effects

The summarized gradient weight $\frac{\partial \mathcal{L}(\tilde{y})}{\partial z_k}$ of existing methods.

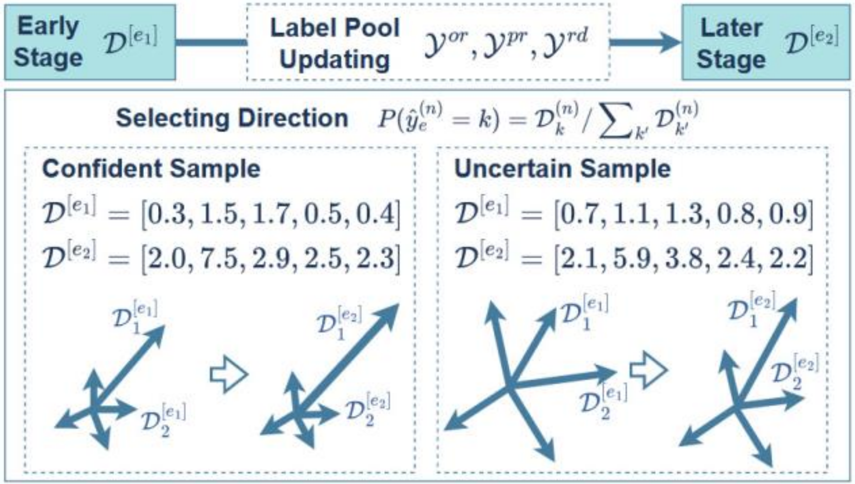
Method	Formula of gradient weight $\partial\mathcal{L}(\tilde{y})/\partial z_k$
Cross Entropy	$p_k - q_k$
GCE [16]	$p_{\tilde{y}}^{\gamma}(p_k - q_k)$
SL [13]	$(\alpha + \beta\lvert A\rvert p_y)(p_k - q_k)$
ELR [8]	$(p_k - q_k) + \frac{\sum_i p_i \hat{p}_i - \hat{p}_k}{1-\sum_i p_i \hat{p}_i}\theta p_k$
Peer Loss [10]	$(p_k^{(n)} - q_k^{(n)}) - (p_k^{(n1)} - q_k^{(n2)})$
EG Reweighting [11]	$w_{EG}\,(p_k - q_k)$
CIW [4]	$w_{CIW}\,(p_k - q_k)$
Co-teaching [2]	$\mathbb{1}\,[\mathcal{L}(p^*)_y < \tau']\,(p_k - q_k)$
DivideMix [5]	$\mathbb{1}\,[GMM\,(\mathcal{L}(p^*)_y) > \tau'']\,(p_k - q_k)$

Existing methods essentially enhance or inhibit the gradient weight term $\frac{\partial \mathcal{L}(\tilde{y})}{\partial z_k}$.

- Samples with low confidence would be reduced or removed to avoid the influence of noise but therefore cannot be exploited in model training.
- Methods with semi-supervised learning train uncertain samples based on unreliable predictions, new noise will be introduced on another fixed direction.

Gradient Switching Strategy (GSS)

Instead of switching the gradient into another fixed direction,
GSS is proposed to select directions with dynamic probabilities.



The updating of the gradient direction pool is based on three strategies:

Original: $v^{or} = p_{\tilde{y}}(1 - e/E),$

Predicted: $v^{pr} = p_{\tilde{y}}(\lambda_1 e/E),$

Random: $v^{rd} = \lambda_2 e/E,$

- Mislabeled but easy samples will be highly confident and generate explicit principal directions. Thus these samples can be trained in correct directions, rather than being misled by original labels or removed directly.
- For uncertain samples, the gradients switch more randomly across all categories, which allows the model to explore in various directions without being affected by the continuous damage. Principal directions can be generated as the model performance improves.

Experiment

The gradient bias of each sample with the noisy label $\tilde{y}$ and clean label y:

$$\Delta g = \sum_{e}^{\mathcal{E}} \mu a^e \otimes \left| d_{\tilde{y}}^e - d_y^e \right|$$

The experimental analysis of various methods' gradient biases in different training stages:

	Dataset	CIFAR-10			CIFAR-100		
	Epochs	50	100	150	50	100	150
Gradient Bias ($\times 10^2$)	Δg_{ori}	2.15	5.65	17.62	5.26	13.26	26.91
	Δg_{sc}	1.22	2.70	6.67	3.36	7.31	14.52
	Δg_{ssl}	**1.18**	2.64	6.71	3.29	7.20	18.34
	Δg_{gss}	1.20	**2.61**	**6.53**	**3.27**	**7.04**	**12.19**

Original: $\Delta g_{ori} = \mu a \otimes \left| \mathcal{E} \left(d_{\tilde{y}} - d_y \right) \right|$

Sample Screening: $\Delta g_{sc} = \mu a \otimes \left| \mathcal{E} \left(-\sum_k \frac{\partial \mathcal{L}(y)}{\partial z_k} d_k \right) \right|$

$$= \mu a \otimes \left| \mathcal{E} \left(\sum_{k \neq y} p_k d_k - (1 - p_y) d_y \right) \right|$$

Semi-supervised Learning: $\Delta g_{ssl} = \mu a \otimes \left| \sum_k \left(\left(\frac{\partial \mathcal{L}_{ssl}}{\partial z_k} - \frac{\partial \mathcal{L}(\tilde{y})}{\partial z_k} \right) \frac{\partial z_k}{\partial m^l} \right) \right|$

GSS (ours): $\Delta g_{gss} = \mu a \otimes \left| \sum_{e}^{\mathcal{E}} \left(d_{\hat{y}_e} - d_y \right) \right|$

- In methods of sample screening, the filtered samples can not be used in training, which causes bias with clean labels.
- SSL has a relatively low bias at the early stage, but the bias increases more compared to Δg_{sc} and Δg_{gss}, which might be due to the added noise by using predictions as targets for mislabeled samples.

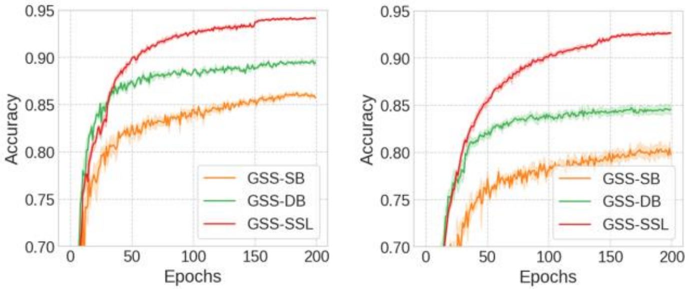
Experiment

Dataset	Method \Ratio	Symmetric				Asymmetric		
		20%	40%	60%	80%	20%	30%	40%
CIFAR-10	GCE [51]	88.77±0.18	84.66±0.30	78.43±0.25	66.11±0.27	87.28±0.13	84.63±0.15	82.15±0.27
	SL [41]	88.98±0.20	84.62±0.28	78.22±0.25	68.53±0.26	84.94±0.19	80.90±0.22	78.71±0.21
	ELR+ [22]	87.77±0.30	83.87±0.28	79.19±0.30	62.01±0.32	84.35±0.20	82.36±0.22	80.56±0.29
	Co-teaching [10]	89.59±0.09	87.20±0.20	81.40±0.15	72.94±0.21	85.99±0.12	84.23±0.11	79.48±0.12
	JoCoR [50]	86.82±0.24	85.31±0.22	76.50±0.23	66.94±0.33	86.73±0.18	79.84±0.17	77.19±0.24
	DivideMix [19]	94.26±0.14	92.85±0.19	92.26±0.21	90.07±0.17	92.98±0.15	91.57±0.13	90.59±0.16
	GSS-SSL (Ours)	**94.31±0.12**	**94.20±0.11**	**92.84±0.25**	**91.61±0.21**	**93.42±0.10**	**92.44±0.12**	**91.82±0.10**
CIFAR-100	GCE	69.19±0.24	63.17±0.35	52.45±0.32	22.60±0.40	67.19±0.30	55.41±0.28	49.75±0.28
	SL	70.43±0.29	62.28±0.31	53.20±0.45	25.79±0.42	69.11±0.28	57.63±0.30	52.06±0.27
	ELR+	66.77±0.33	63.89±0.26	49.93±0.26	19.81±0.33	64.10±0.28	51.89±0.36	46.78±0.35
	Co-teaching	70.35±0.19	64.54±0.20	52.99±0.22	27.05±0.24	69.96±0.23	58.84±0.39	55.74±0.35
	JoCoR	65.36±0.27	61.70±0.24	50.33±0.31	18.44±0.40	64.01±0.41	53.40±0.49	48.99±0.48
	DivideMix	75.89±0.14	73.90±0.16	67.41±0.16	45.82±0.15	72.20±0.20	69.04±0.19	59.16±0.19
	GSS-SSL (Ours)	**76.71±0.19**	**76.10±0.20**	**71.92±0.21**	**55.04±0.25**	**73.81±0.22**	**72.20±0.27**	**65.84±0.20**

Classification results on CIFAR-10 and CIFAR-100 with different ratios of symmetric/asymmetric noise.

Method	Clothing1M	WebVision		ILSVRC12	
	Top1	Top1	Top5	Top1	Top5
GCE [51]	71.73	61.22	80.81	59.13	79.09
SL [41]	72.05	63.78	84.29	61.56	84.08
ELR+ [22]	71.48	63.61	83.50	60.10	83.13
Co-teaching [10]	72.50	64.09	85.01	62.94	84.76
JoCoR [50]	71.74	60.79	82.48	57.15	81.33
DivideMix [19]	74.59	77.21	91.60	**75.23**	90.76
GSS-SSL (Ours)	**74.88**	**77.35**	**93.09**	75.18	**92.84**

Classification results on real-world noisy datasets.



(a) 40% Noisy labels (b) 60% Noisy labels

The ablation results of GSS combinations with various frameworks.

Conclusion

- This paper makes a deep analysis from a new perspective of gradient directions, demonstrating that label noise can cause continuous damage throughout the model training.
- The Gradient Switching Strategy (GSS) is proposed to prevent the continuous gradient damage of mislabeled samples to the model training.
- Detailed theoretical analysis and extensive experimental results show that the proposed GSS can effectively prevent damage of mislabeled samples.

Thanks for Listening